Conceptos sobre minería web

AGUILAR JOSÉ
Doctor en Infromatica Director
CEMISID-ULA. Dpto. de Computación
Facultad de Ingeniería. Universidad de los Andes.
aguilar@ing.ula.ve
Mérida-VENEZUELA

ALTAMIRANDA JUNIOR

Especialista en Computación
Universidad de los Andes.
Mérida-VENEZUELA

RESUMEN.

En este trabajo se realiza una introducción a la Minería Web. Se presenta los diferentes usos que se han hecho de ella, las perspectivas derivadas de ese uso, y los posibles avances que se pueden esperar. Previo a esto, se establece el marco teórico de donde se deriva la Minería Web, como lo es la Minería de Datos y los Sistemas de Descubrimiento de Conocimiento. Así, este articulo presenta el estado de arte en el área de Minería Web.

PALABRAS CLAVES
Minería Web, Minería de Datos, Internet, Descubrimiento de Conocimiento

1. INTRODUCCIÓN

Internet es un gran deposito de información que crece constantemente. En ella existe una infinidad de sitios que necesitan ser visitados y clasificados a la hora de hacer una búsqueda. Existen, y son muy conocidas, las poderosas herramientas de búsqueda que tratan de encontrar información por categoría o por contenido, tales como Altavista, Yahoo, Google, etc. A estos buscadores se les introducen palabras claves, y ellos determinan las páginas o sitios Web que contienen dichas palabras, tratando de satisfacer así los requerimientos del usuario. Muchas veces estas consultas traen resultados inconsistentes, o documentos que cumplen con el criterio de búsqueda pero no con el interés del usuario. Por esta razón, para superar este problema, en los últimos años han surgido una serie de técnicas que permiten el procesamiento avanzado de datos sobre la Internet, los cuales realizan un análisis en profundidad de los mismos de forma automática, denominada Minería Web (Web Mining). De esta manera, Minería Web puede definirse como la aplicación de técnicas de Minería de Datos en Internet para el descubrimiento y análisis de páginas Web, la generación de patrones para clasificar la información

de las páginas Web, entre otras cosas, [2, 3, 7, 9].
En este trabajo haremos una presentación de las diferentes técnicas que han sido usadas para realizar Minería en la Web, así como los resultados más exitosos encontrados con ellas, de tal forma de establecer el estado de arte de esta área. El articulo esta organizado de la siguiente forma: la sección 2 hace una introducción a la Minería de Datos, presentando una metodología para el problema de Descubrimiento de Conocimiento. La siguiente sección presenta una caracterización de la Web, para después presentar los diferentes mecanismos de búsqueda en la Web. La sección 4 hace una introducción a la Minería Web, presentando exhaustivamente los diferentes tipos de Minería Web, algunas técnicas de Minería Web, así como sus usos. Finalmente, la sección 5 presenta diferentes aplicaciones de la Minería Web (por ejemplo, en tareas de mercadeo, definición de perfiles de clientes, problemas de seguridad informática, entre otros).

2. MINERÍA DE DATOS

La minería de datos es un término genérico que engloba las técnicas y herramientas usadas para extraer información útil desde grandes bases de datos. Los algoritmos de minería de datos se enmarcan en un proceso completo de extracción de información conocido como "Descubrimiento de Conocimiento en Bases de Datos"-DCBD (sus siglas en ingles son KDD-Knowledge Discovery in Databases), que se encarga, además, de la preparación de los datos y de la interpretación de los resultados obtenidos. En general, las técnicas de minería de datos intentan obtener patrones o modelos a partir de los datos recopilados. Este proceso involucra un análisis de los datos, el reconocimiento de patrones sobre el conjunto de datos y, una clasificación o agrupación de los mismos.

En general, se deben interpretar grandes cantidades de datos y encontrar relaciones o patrones en ellos. Los patrones descubiertos han de ser válidos y potencialmente útiles (ver Fig. 1). Entre las técnicas que pueden ser usadas en tareas de minería de datos tenemos: árboles de decisión, redes neuronales, algoritmos genéticos, búsqueda de asociaciones, programación genética, lógica difusa, etc. Dichas técnicas toman los datos y los transforman en información útil y entendible [2, 6, 9, 10].

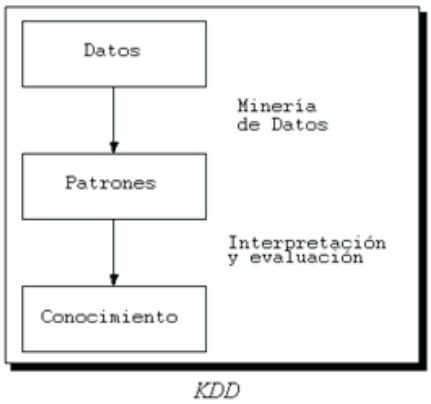

Figura 1. Modelo del Proceso de DCBD [10].

El DCBD es interactivo, consta de varios pasos en los cuales se deben tomar decisiones. Estos pasos son[2, 6, 10]:

- **COMPRENDER EL DOMINIO DE LA APLICACIÓN:** Conocer el dominio del problema es fundamental para realizar el proceso de búsqueda de información. Es necesario conocer qué problema se está tratando y entenderlo, para luego seleccionar los datos que han de ser analizados.
- **PREPARAR EL CONJUNTO DE DATOS:** Se debe realizar una limpieza de los datos, la misma consiste en eliminar datos repetidos o que no tienen sentido en el dominio de la aplicación.
- **ELEGIR EL MODELO A EMPLEAR:** Dependiendo del tipo de problema y el tipo de conocimiento que deseemos obtener, pueden elegirse diferentes modelos, entre esos:
- **CLASIFICACIÓN**: Consiste en distribuir el conjunto de datos en un número dado de categorías.
- **AGRUPACIÓN**: Técnica que permite la identificación de grupos en los cuales los elementos que los conforman guardan una gran similitud entre sí.
- **DESCUBRIR CONOCIMIENTO:** Es la fase donde se obtienen los patrones sobre los datos. Dependiendo del modelo elegido en la fase anterior, se utilizarán distintas técnicas: redes neuronales, algoritmos genéticos, programación genética, sistemas difusos.
- **PREPROCESAR LOS CONOCIMIENTOS OBTENIDOS:** Los patrones obtenidos en la fase anterior se analizan para seleccionar los útiles y ordenarlos según el valor de utilidad, para después interpretarlos.
- **UTILIZAR LOS CONOCIMIENTOS OBTENIDOS:** Todas estas fases tienen como objetivo final ayudar a la toma de decisiones, es decir, se necesitará una persona o un equipo que estudie los resultados obtenidos y use los mejores patrones en los procesos de toma de decisiones.

3. CARACTERIZACIÓN DE LA WEB

La Web es vista como una inmensa y dinámica colección de páginas que incluyen innumerables enlaces e inmensos volúmenes de accesos. Ella provee una rica y nunca vista fuente de datos. Sin embargo, la Web genera varios retos para que sea un recurso eficaz y pueda descubrir conocimiento [1, 3, 4, 5, 7, 8, 9, 11]:

- La complejidad de las páginas Web excede la complejidad de cualquier colección de documentos de texto tradicional. Aunque la Web funciona como una enorme biblioteca digital, las páginas, en si mismas, carecen de una estructura uniforme y contienen muchos más estilos y contenidos variados que cualquier conjunto de libros o documentos basados en texto tradicional. Por otra parte, el gran número de documentos que esta biblioteca digital contiene, no pueden ser indexados, lo cual hace que la búsqueda de los datos en sus contenidos sea extremadamente complicado.
- La Web constituye una inmensa fuente de

información dinámica. La Web no sólo crece continúa y rápidamente, la información que contiene también recibe actualizaciones constantes. Los enlaces a la información y los registros de acceso también sufren frecuentemente actualizaciones.

- La Web sirve a un gran número de usuarios. La Internet se ha extendiendo rápidamente y los usuarios están conectando a través de millones de estaciones de trabajo. Estos usuarios tienen intereses y propósitos diferentes de uso de la Web. Muchos carecen del conocimiento de la estructura de la red de información, desconocen el costo de la búsqueda, frecuentemente se pierden dentro del océano de información de la Web, y pueden acceder a muchas páginas o pasar por largas esperas antes de obtener los resultados deseados.

Sólo una pequeña porción de las páginas Web contiene la información verdaderamente relevante o útil. Un usuario dado generalmente se enfoca en sólo una porción diminuta de la Web, ignorando el resto de los datos.

3.1 MECANISMOS DE BÚSQUEDAS EN LA WEB

En general, los usuarios pueden escoger entre tres grandes métodos para acceder a la información almacenada en la Web [1, 7, 9]:

1. Búsqueda a través de palabras claves, usando herramientas como Google o Yahoo, las cuales usan las palabras claves para encontrar documentos específicos.
2. Preguntando exhaustivamente a la página Web donde esta la fuente de información. Por ejemplo, preguntar por los datos de un libro en amazon.com, o los datos inmobiliarios en realtor.com. Esta información esta escondida en la base de datos que está detrás de una página Web dada, y se accede a ella a través de un formato de preguntas, de lo contrario no podría accederse.
3. Navegando aleatoriamente siguiendo los enlaces de la Web.

Sin embargo, los mecanismos de búsqueda actuales tienen varias deficiencias [3, 7, 9]:

1. Carencia de alta calidad en la búsqueda por palabras claves, por ejemplo:
   a) Una búsqueda devuelve a menudo muchas respuestas, sobre todo si las palabras claves propuestas incluyen términos que pertenecen a categorías populares como deportes, política o entretenimiento.
   b) Al usar una palabra clave con mucho valor semántico, se pueden devolver muchas respuestas de baja calidad, por ejemplo, dependiendo del contexto, la palabra clave "jaguar" podría ser un animal, un automóvil, un equipo deportivo o una computadora.
   c) Una búsqueda puede devolver muchas páginas que no contienen explícitamente la palabra clave.
2. Carencia de un eficaz acceso en la Web. Las bases de datos que contiene la Web proporcionan información de alta calidad, actualizada, pero no son fácilmente accesibles, porque los buscadores actuales no pueden preguntar directamente a las bases de datos. Los datos que estas contienen, en la mayoría de los casos, permanecen invisibles a los mecanismos de búsqueda tradicionales. Es decir, la Web proporciona una colección sumamente grande de bases de datos autónomas y heterogéneas, cada una con sus propios sistemas de consultas.
3. Carencia de directorios construidos automáticamente. Un tema o tipo orientado de directorio de información Web presenta un cuadro organizativo de un sector de la Web y soporta búsquedas eficientes de la información de ese sector. Por ejemplo, el siguiente enlace jerárquico "Venezuela > Universidades> Ciencia de la Computación> Programa de Estudio", haría la búsqueda más eficiente. Desafortunadamente, los diseñadores de páginas Web deben construir los directorios, los cuales quedan pre-definidos, por lo tanto, estos proporcionan sólo alcances de búsqueda limitados.
4. Carencia de semántica en las preguntas. Los mecanismos de búsqueda proporcionan pocas opciones para las posibles combinaciones de las palabras claves, por ejemplo: "con todas las palabras", "con cualquiera de las palabras". Algunos servicios de búsqueda en la Web, como Google y Yahoo, proporcionan algunas formas de búsquedas avanzadas, incluyendo "con las frases exactas," "sin ciertas palabras," y con las restricciones en la fecha y tipo de dominio de la página.
5 Carencia de realimentación humana en las tareas de concepción de páginas Web. Los diseñadores de páginas Web proporcionan los enlaces a otras páginas, y nos obligan a recorrer aquéllas páginas Web que ellos encuentran más interesantes o de gran calidad. Infortunadamente, como las actividades humanas y los intereses cambian con el tiempo, los enlaces de la Web puede que no se actualicen para reflejar estas tendencias.
6 Carencia de análisis. Las búsquedas en la Web actuales confían en los índices basados en las palabras claves, y no en los datos reales que las páginas Web contienen, los mecanismos de búsqueda sólo proporcionan un limitado apoyo para el análisis de la información de la Web.

4. MINERIA WEB

Minería Web puede definirse como el análisis automático y el descubrimiento de información útil desde documentos y servicios de la Web. La Minería Web se basa en la aplicación de algoritmos de minería de datos sobre la información que se encuentran en Internet. La Minería Web se puede descomponer en [3, 7, 9]:

- Descubrimiento de la información: consiste en recuperar documentos Web.
- Selección de la información y pre-procesamiento: consiste en seleccionar y pre-procesar la información obtenida del paso anterior.
- Generalización: consiste en descubrir patrones generales desde los sitios Web visitados.
- Análisis: consiste en validar y/o interpretar los patrones encontrados.

4.1 TIPOS DE MINERIA WEB

Como se infiere de lo anterior, las técnicas de Minería Web pueden ser utilizadas para acceder de manera más eficiente a la información contenida en la Web, de forma directa o indirectamente. Dentro de esta amplia definición existen tres

formas de uso de la Minería Web: minería sobre el contenido en la Web, minería sobre las estructuras de la Web, y minería sobre el uso de la Web [2, 3, 4, 5, 7, 8, 11].

4.1.1. Minería sobre el contenido en la Web

La Minería sobre el Contenido en la Web se refiere a la búsqueda automática de información y extracción de conocimiento a partir del contenido de la página Web, y de las descripciones de documentos en la Web. La heterogeneidad y la estructura variada de las fuentes de información en la Web, hace complicado el descubrimiento automático, organización y manejo de la información. Por ello, los mecanismos de búsqueda actuales basados en palabras claves, tales como Google, Yahoo, Altavista; tienen varias deficiencias. Así, una palabra cualquiera puede ser contenida fácilmente en centenares de miles de documentos. Esto puede llevar a un mecanismo de búsqueda a devolver un número grande de enlaces a documentos, muchos de los cuales son marginalmente relevantes al tema, o contienen sólo material de pobre calidad. Pero, peor aún, muchos de los documentos relevantes puede que no contenga las palabras claves que explícitamente definen el tema. Por lo tanto, estas herramientas no proporcionan información estructurada, ni categorizan, filtran o interpretan documentos.

Estos factores han motivado a investigadores a desarrollar herramientas más inteligentes para la recuperación de información, tales como agentes de búsqueda inteligente, así como también, proveer técnicas de minería de datos para reforzar la calidad de las búsquedas sobre la Web. Algunos de estos esfuerzos son [2, 3, 6, 7]:

- **AGENTES DE BÚSQUEDA INTELIGENTE:** Agentes de búsqueda en la Web basados en inteligencia artificial, que pueden actuar autónomamente o semi-autónomamente, descubriendo y organizando información de la Web. Ejemplos de estos agentes son los motores o "robots" de búsqueda, que permiten encontrar documentos e información en la Web.
- **BASES DE DATOS DE LA WEB:** La idea es desarrollar técnicas para integrar y organizar la información heterogénea y poco estructurada de la Web en colecciones de datos más estructuradas, tales como bases de datos relacionales, y usar métodos estándar de consulta y técnicas de minería de datos para acceder y analizar esta información.
- **TÉCNICAS DE MINERIA DE DATOS USADAS EN LA BUSQUEDA:** Se utilizan para descubrir información en la Web. Por ejemplo, un mecanismo de búsqueda basado en minería de datos podría investigar el conjunto de documentos de la Web obtenidos usando esquemas tradicionales de búsqueda, para seleccionar un conjunto más pequeño de documentos realmente importantes para el usuario.

4.1.2. Minería sobre las estructuras de la Web

La Minería sobre las Estructuras de la Web se refiere al proceso de inferir conocimiento a partir de las referencias o enlaces entre documentos de la Web. Por ejemplo, muchos enlaces que apuntan a un documento pueden indicar la popularidad de un documento (páginas autoritarias), mientras los enlaces que salen de un documento pueden indicar la riqueza o variedad de temas que
cubre el documento [4, 5, 7].

Así, la importancia de una página Web está en los enlaces. Cuando un diseñador de una página Web crea un hipervínculo a otra página Web, esta acción puede ser considerada como una transferencia a esa página. El enlace colectivo a una página dada por diseñadores diferentes, puede indicar la importancia de la página y pueda llevar naturalmente al descubrimiento de páginas Web "claves". Así, los enlaces en la Web proporcionan una rica fuente de minería de datos. Pero se debe tener en cuenta los siguientes aspectos [4, 7, 9]:

1. No todos los hipervínculos de una página representan la transferencia a una búsqueda que se está realizando. Los diseñadores de páginas Web crean algunos enlaces para otros propósitos, tales como la navegación a anuncios pagados.
2. Una página que pertenece a un producto o sobre un asunto en particular, raramente tendrá en ella un enlace a la competencia rival. Por ejemplo, La Coca-Cola evitará crear un hipervínculo a la Pepsi, y ésta otra se asegurará que ningún enlace a la Coca-Cola aparezca en la página Web de Pepsi.
3. Las páginas "claves" raramente proporcionan descripciones claras de ellas. Por ejemplo, la página Web de Yahoo no puede contener la descripción explícita de sí misma.

Estas propiedades de la estructura de los enlaces Web han llevado a los investigadores a considerar categorías importantes de páginas Web: los cubos. Un cubo es una sola página Web, o conjuntos de páginas que proporcionan enlaces a páginas "claves". Un cubo implícitamente confiere el estado de autoridad o clave a los sitios que enfocan un tema específico. Generalmente, un buen cubo apunta a muchas páginas autoridades o claves, y, recíprocamente, una página que es apuntada por muchos cubos puede ser considerada una buena autoridad. A partir de la relación de refuerzo mutuo entre los cubos y las páginas autoridades, se pueden usar técnicas de minería de datos para automatizar el descubrimiento de estructuras y recursos Web de alta calidad [7, 9].

4.1.3. Minería sobre el uso de la Web

La Minería sobre el uso de la Web es un tipo de Minería Web que se refiere al descubrimiento y análisis de los patrones de accesos o hábitos de los usuarios, los cuales se extraen desde la información implícita de sus actividades [3, 5, 8, 11]. Las organizaciones y compañías generan y almacenan grandes volúmenes de datos en sus funcionamientos diarios sobre Internet. La mayoría de esta información es generada automáticamente por los servidores Web y se almacenan en archivos llamados "log files" en esos servidores. Otras fuentes de información sobre el usuario incluyen las referencias de sus páginas Web a otros sitios o páginas Web, y sus registros en bases de datos vía formularios en línea. Analizar tales datos puede ayudar a las organizaciones a determinar que usuarios visitan su sitio, permitiendo generar estrategias de mercadeo de productos y aumentar la efectividad de sus campañas promociónales, entre otras cosas. El análisis del acceso al servidor y los datos de registro del usuario también puede

proporcionar información valiosa de cómo mejorar la estructura del sitio, creando una presencia en Internet más eficaz para las organizaciones.

Para llevar a cabo un proceso de Minería sobre el uso de la Web, es necesario realizar el siguiente procedimiento: En una primera fase, debemos establecer los objetivos que se persiguen alcanzar en la tarea de Minería Web, así como las estrategias de validación de estos objetivos. En la siguiente fase se reúnen los datos que formarán parte del análisis, pudiendo venir de archivos históricos de "logs files" del servidor o servidores del sitio Web a analizar, de datos de los clientes/usuarios, de datos de facturación, etc. Una vez recopilados todos los datos, se llevarán a cabo tareas de limpieza y selección de los mismos, donde se identificarán las sesiones y transacciones del usuario [8, 9]. Después se realiza la integración de los datos. Como resultado de esta fase se construirán los archivos o almacenes de datos sobre los que se le aplicaran las diferentes herramientas de extracción de información. Finalmente, se realizará las tareas propiamente de análisis sobre dichos datos (descubrimiento de patrones, análisis de los comportamientos de los usuarios, etc.).

Debido a la existencia de cachés en distintos niveles de la conexión del cliente/usuario con el servidor Web, algunas páginas que el cliente/usuario recibe no quedan registradas en los "log files", esto dificulta el proceso de minería. Para resolver este problema hay distintas posibilidades: por ejemplo, se puede reconstruir la secuencia real de páginas visitadas a partir de los rastros que queden en el fichero log y en el mapa de recorrido de los servidores Web [8, 9].

4.2 ALGUNAS TÉCNICAS DE MINERÍA WEB

Las técnicas que más se emplean para realizar Minería Web son [2, 3, 4, 7, 8, 9]: agrupamiento, clasificación, detección de reglas de asociación, análisis de caminos, detección de patrones secuenciales, entre otras.

- **AGRUPAMIENTO Y CLASIFICACIÓN:** Las técnicas de agrupamiento o clustering distribuyen comportamientos de individuos similares en grupos homogéneos, o de información semejante en grupos. Por ejemplo, dependiendo de la información almacenada en los "files log", es posible detectar grupos de usuarios como [5, 9]:

    1.Aquellos que visitan gran cantidad de páginas con un tiempo de estancia muy similar en todas ellas.

    2.Los que visitan un número pequeño de páginas en sesiones cortas.

Al tener descubiertos los prototipos o perfiles de cada grupo, se pueden usar las características de cada uno de ellos para realizar tareas de clasificación. Las técnicas de clasificación permiten desarrollar un perfil para clientes/usuarios que acceden a archivos concretos del servidor, en función de sus patrones de acceso o de la información extraída. El agrupamiento y/o clasificación de clientes/usuarios o de la información, puede facilitar el desarrollo y la ejecución de estrategias futuras, tales como envío de correo automático a aquellos clientes/usuarios que se encuentren dentro de un cierto grupo, o presentación de contenidos específicos según el tipo de grupo, entre otras cosas. Todo esto es hecho a través de la minería de Web [9].

- **REGLAS DE ASOCIACIÓN:** El problema consiste en descubrir todas las asociaciones y correlaciones entre los accesos y usos de la Web por parte de los usuarios. Cada sesión o transacción consiste en un conjunto de URL's accedidas por un cliente en una visita al servidor. Con estas técnicas podemos encontrar correlaciones tales como: [5, 9]:

    1. El 40% de clientes/usuarios que accedieron la página Web con URL /entidad/productos/producto1.html, también accedieron a la página Web /entidad/productos/producto2.html;
    2. El 30% de clientes/usuarios que accedieron a /entidad/anuncio/oferta-especial.html, efectuaron un pedido interactivo en /entidad/productos/producto1.

Las reglas de asociación descubiertas a partir de los históricos de acceso Web, dan una indicación de cómo disponer mejor el espacio Web de una organización. Por ejemplo, si se descubre que el 80% de los clientes/usuarios que acceden a entidad/productos y a /entidad/productos/página1.html también acceden a entidad/productos/página2.html, pero sólo el 30% de aquellos que accedieron a /entidad/productos también accedieron a la página /entidad/productos/página2.html, entonces es probable que cierta información en página1.html lleve a los clientes/usuarios a acceder a página2.html. Esta correlación podría sugerir que esta información debería moverse a un nivel más alto (por ejemplo, /entidad/productos), para incrementar el acceso a página2.html.

- **ANÁLISIS DE CAMINOS:** Esta técnica supone la generación de algunas formas de grafos orientados que representan relaciones entre páginas Web. Este grafo puede ser un esquema físico en el que las páginas Web son los nodos del grafo y los hiper-enlaces entre las páginas son las flechas dirigidas entre nodos. Pueden formarse otros grafos a partir de los tipos de páginas Web, con arcos que representen la similitud entre páginas, o creando arcos que muestren el número de usuarios que van desde una página a otra. Ejemplos de información que puede descubrirse a partir de un análisis de camino son [4, 5, 7]:

    1. El 70% de clientes/usuarios que accedieron a /entidad/productos/página2.html lo hicieron a partir de /entidad siguiendo por /entidad/productos, y /entidad/productos/página1.html. Esta regla sugiere que hay información útil en las páginas visitadas hasta llegar a /entidad/productos/página2.html, ya que los usuarios navegaron a través de ellas para llegar a pagina2.html.
    2. El 80% de clientes/usuarios que accedieron al lugar empezaron por /entidad/productos. Esta regla afirma que la mayor parte de los usuarios están accediendo al sitio a través de una página diferente de la página principal (/entidad en este ejemplo).
    3. El 65% de clientes/usuarios abandonaron la búsqueda en Internet después de consultar cuatro o menos páginas. Esta regla indica que la mayoría de usuarios no navegan más de cuatro páginas, sería adecuado

asegurarse de que la información más importante está contenida dentro de las cuatro páginas más cercanas a los puntos de entrada.

Así, el análisis de caminos podría utilizarse para determinar los caminos más frecuentemente seguidos en un sitio Web.

• **PATRONES SECUENCIALES:** El problema de descubrir patrones secuenciales se centra en localizar la presencia de un conjunto de elementos seguidos por otro elemento en un conjunto de transacciones o visitas ordenadas en el tiempo. En un histórico de transacciones de un servidor Web, la visita de un cliente se guarda por un periodo de tiempo. El descubrimiento de patrones secuenciales en los históricos de acceso al servidor Web permite predecir los patrones de visita de los usuarios. Analizando esta información, se pueden determinar relaciones temporales entre elementos de datos, tales como [4, 5, 7, 9]:

1. El 30% de clientes/usuarios que visitaron /entidad/productos/, habían hecho una búsqueda en Yahoo la semana anterior con la palabra clave w.
2. El 60% de clientes/usuarios que efectuaron una orden de compra on-line en /entidad/productos/producto1.html, también efectuaron una orden on-line en /entidad/productos/producto4.htm con un lapso máximo de 15 días.

Otro tipo de dependencia de datos son las secuencias ocurridas en un intervalo dado. Por ejemplo, podemos estar interesados en encontrar las características comunes de todos los clientes/usuarios que visitaron una página en particular dentro del periodo de tiempo [t1,t2], o puede que estemos interesados en un intervalo de tiempo (un día, una semana, etc.) en el cual una página en concreto es la más accedida [9].

En general, con la información obtenida con estas técnicas, se puede responder a preguntas tales como:

- ¿Qué tipo de visitantes navega por un sitio Web dado?
- ¿Qué tipo de visitantes prefiere un determinado contenido?
- ¿Qué rasgos o características tienen en común los visitantes de una página Web dada?
- ¿Quiénes son los usuarios más fieles a una página Web dada?
- ¿Qué contenido y estructura es la más aceptada por mis usuarios?

5. BENEFICIOS DE LA MINERIA WEB

5.1 PERSONALIZACIÓN

La Minería Web permite a los proveedores de servicios Web influenciar a sus clientes mediante el entendimiento y predicción de su comportamiento. Los beneficios comerciales de la Minería Web permiten a los proveedores entregar servicios personalizados, definir su estrategia de productos y servicios, detectar fraudes, etc. En resumen, les da la habilidad de servir las necesidades de sus clientes y entregarles el mejor y el más apropiado servicio en un momento dado [1, 3, 11].
Por otro lado, la Minería Web juega un rol importante en el área de mercadeo. Mediante este análisis puede entregarse mensajes personalizados a las personas individualmente. La oportunidad de identificar a los navegantes y dirigir a los compradores de productos y servicios a ofertas atractivas y promociones de venta, es sin duda muy interesante [1, 3, 5, 11].

Así, la Minería Web permite la personalización de la Web, tal que un sitio Web pueda ajustarse de acuerdo a las preferencias y perfil de cada usuario en particular. Esto permite a los proveedores de servicios Web desarrollar relaciones fieles y duraderas con cada individuo o visitante.

5.2 ENTENDIMIENTO LA CONDUCTA DEL CONSUMIDOR

Los proveedores de servicios Web pueden obtener conocimiento de los gustos y preferencias de los visitantes de su sitio. Con ello, entre otras cosas pueden [3, 5, 9]:

- Descubrir y comparar modelos de usuarios.
- Aprender quién está accesando a su sitio.

5.3 DETERMINAR LA EFECTIVIDAD DEL SITIO WEB

Con la Minería Web, las compañías pueden descubrir las áreas de alto y bajo impacto de su sitio Web. Los administradores del sitio Web ya no tienen que confiar en la intuición al diseñar un esquema del sitio [5, 8 11].

5.4 MEDICIÓN DEL ÉXITO DE MERCADEO

En el mundo físico es difícil medir el éxito de las campañas de comercialización. Por el contrario, en Internet se pueden obtener medidas reales del éxito de una campaña de mercadeo [1, 5, 8]. Usando Minería Web las compañías pueden generar modelos para describir el éxito o no de sus campañas de mercadeo, y en base a eso hacer los ajustes en sus sitios Web.

5.5 SEGURIDAD

La Minería Web permite la detección de accesos inusuales a datos privados. Así, la Minería Web hace que las páginas y el comercio electrónico sean más seguros a los ataques de piratas cibernéticos o al acceso a la información por usuarios no autorizados [3, 7].

5.6 ANÁLISIS DE TRÁFICO DE REDES

La Minería Web permite la determinación de los requerimientos de equipo y la distribución de datos con el fin de manejar eficientemente el tráfico de un sitio [3, 7, 9].

6. CONCLUSIONES

1. La Minería Web es un proceso que permite la aplicación de las técnicas de Minería de Datos para extraer patrones e información útil de los usuarios de la Web, tales como: correlaciones entre las páginas Web y grupos de usuarios, comportamientos de los usuarios al navegar por Internet, agrupamiento de páginas según los usuarios, entre otras cosas. Esta tecnología, utilizada inicialmente en bases de datos convencionales, tiene en la Web unas aplicaciones potenciales e inmediatas, que apenas están iniciándose.
2. La Minería Web puede aportar información valiosa,

tanto a los gestores de servicios de información, a los proveedores de servicios de Internet, como a los buscadores de información sobre la Internet.

3. La Minería Web pone a disposición de los proveedores de servicios Web un conjunto de herramientas para entender y predecir la conducta de sus usuarios. Las instituciones/compañías pueden ahora perfeccionar sus sitios para tener el máximo impacto y personalizar el contenido de su sitio Web.

4. La Minería Web permite conocer las necesidades especificas de cada usuario de Internet. Esta información puede ser analizada en asociación con el contenido del sitio Web, para lograr atraer a otros usuarios o hacer modificaciones según los requerimientos de los usuarios.

5. La Minería Web apenas esta empezando a ser usada en Internet. En el futuro, los procesos de búsqueda en Internet serán rotundamente mejorados con esta herramienta, aprovechando el contenido semántico creado por dichas técnicas.

7. REFERENCIAS

[1] Aguilar J., Velásquez, L., Pool. M. "La Web como plataforma tecnológica para soportar aplicaciones de inteligencia de negocios", Universidad, Ciencia y Tecnología, Vol. 7, No 27, pp. 169-178, 2003.

[2] Aguilar J., Altamiranda, J. "Minería de Datos en la Web usando Computación Evolutiva", Ingeniería de Software en la Década del 2000s. (Ed. N. Brisaboa), AECI, RISTOS2, España, pp. 153-168, 2003.

[3] Aguilar J., Leiss E., Callaos N. "Introduction to Web Computing",, International Institute of Informatics and Systemics, USA, 2003.

[4] Chakrabarti, S. Dom, B. Kumar, R. Raghavan, P. Rajagopalan, S. Tomkins, A. Gibson, D. Kleinberg, J. "Mining the Web's Link Structures" IEEE, pp. 60-66, 1999.

[5] González, J. "Minería Web: Como conocer a nuestros clientes del canal de Internet", XXIV Taller de Ingeniería de Sistemas. Chile. Julio 2001.

[6] Han, J. Chen Chuan, K. "Data Mining for Web Intelligence" IEEE, pp. 54-60, 2002.

[7] Kosala, R. Brockeel, H. "Web Mining Research: A Survey" SIGKDD Exploration, Volume 2, Issue 1, pp. 25-37, 2000.

[8] Masand, B. Zaiane, O. Srivastava, J. Spiliopoulou, M. "Web Mining for usage Patterns & Profiles" SIGKDD Explorations, Volume 4, Issue 2, pp. 125-127, 2002.

[9] "Minería Web: Documento Básico", Múltiples autores, Daedalus, 2002, España, http://www.daedalus.com (actualizada 2002).

[10] Morales, E. "Introducción a la Minería de Datos", Reporte Técnico, ITESM, México, 2001.

[11] Silva, M. "Personalización Inteligente de sitios Web usando Web Mining" XXIV Taller de Ingeniería de Sistemas. Chile. Julio 2001.